Lab Project Part 1 Image Classification using Bag-of-Words

Daniel van de Pavert, Damiaan Reijnaers, Shuai Wang

October 2020

Abstract

Bags of words is a traditional method for class image classification which has the process of feature extraction and description, building a visual vocabulary, quantifying features using visual dictionary and representing images by frequencies of visual words, finally training the classier. In this report, each part will be discussed and realised. There are five classes of image with totally two hundred adn fifty images for training. Another fifty images are used for test. Mean Average Precision is the way of qualitative analysis in this report. We will discuss the classification quality between different size of vocabulary and different ways of getting descriptor(SIFT,RGB-SIFT,SURF,HOG)

1 Introduction

. Bag of words is a method primarily seen in Natural language processing (NLP) but it also has applications in image classification. In contrast with the convolutional neural network(CNN) used in the second part of this assignment, bag of words is a relatively old fashioned method, however does not impede its usefulness as it is still a widely applied method. The application of the Bag of words method in NLP counts the frequency of words in a text, these frequencies are then used to create a histogram of all the words in the text. In image classification area, instead of using words, we count the frequency of image feature occurrence, which we use certain pattern find in images.

The general idea of bag of words or bag of visual words is to represent a image by different features. These features consists of keypoints and corresponding descriptors. This method is invariant to illumination, rotation, translation and scaling. The keypoints are "interest points" within the image and can be used to derive a descriptor. A descriptor is a vector summarizing the properties of the corresponding keypoint. Using the descriptors the keypoints can be classified. In bag of words, we use keypoints and descriptors to build "vocabularies" such that images can be represented by a frequency histogram of images features. With this feature frequency histogram we can categorize images and classify new images. In the following parts of this assignment we will introduce our implementation of Bag of words for usage in image classification.

2 Bags of words theory

As mentioned before, the first step is to detect features and descriptors from each image in the training set. There are plenty of methods to find features like DOG, Harris detector and SIFT. In this report, we use SIFT algorithm for feature extraction. SIFT is the algorithm developed by David Lowe and this part has been explained detailed in last lab so we will not go deep this time. The result of keypoint is as Fig1.

Then we have to Build Visual Vocabulary, and a important concept is Clustering. Clustering is a method of grouping a set of objects in a certain way that objects in same group are more similar than to those in other sets. For realising the cluster, there are some models like Kmeans or Density

Figure 1: circles with size of keypoint).

based clustering. In this project, we choose Kmeans clustering, Suppose there are X objects and to be divided into K clusters, the input can be a set of features, X=x1,x2,....,xn, the goal is basically to find the shortest distance among each point assigned centroid and scatter cloud. We can see that for each cluster centroid, there exists a group of point around it, known as the center. For the initial part, we define an initial random solution, which is cluster centroids. We randomly place them within bounds of data. Then iterating process contains two parts, the first is "assignment step", which is assigning each observation to the cluster with nearest mean and the second part is "update step", Once we are able to realise the first step, most crude clustering, we shall relocate the cluster centroids.The newly calculated cluster centrods can be said to be the aggregate of all member of this cluster.Once the averaging step is accomplished, and the new clusters are computed, the same process is repeated over and over again. This iteration process is going until new cluster nearly same with the last computed cluster.

The next part is to link vocabulary and clustering to build the Using SIFT, we detect and compute features of each image. SIFT returns us a m*128 dimension array, where m is the number of features been extracted. Hence if we train n images, we shall obtain

$$\begin{bmatrix} features_0 \\ features_1 \\ features_2 \\ .... \\ .... \\ features_n \end{bmatrix}$$

where $features_i$ is a array of dimension m*128. After getting all features, we have to group similar features. Now we have a list of visual words using in every image. The next step we group similar features. Since similar features provide an approximate estimate to what the image is, like synonyms tend to express upon the gist of a sentence. Hence, when the system is trained over several images, similar features which describe similar portions of the image are grouped together to develop a vast vocabulary base. Each of these group collectively represent the word and all these groups gives the complete vocabulary getting from training data we put. We simply refer them by signing group number to give the definition.

Encomprising of the total number of each type of feature/word present in the training set in totality. Then we rebuild the histograms by subtracting the mean of the features and dividing them by the standard deviation of the training data.Fig2 shows how a collective vocabulary will look like in each class with SIFT. We can find they have huge difference of frequency of each cluster and this is showing they represent different classes. They also have similarities in certain position so this is the area which may induce classificatin error later. Finally, we can train the Support Vector Machines (SVM). Support Vector Machine classifier builds a hyper plane or set of hyper planes in a high dimensional space that is used for classification. The reason SVM is a popular way used in BoVW models is because its superiority in performance over K Nearest Neighbour(KNN) classification. Hence,we use SVM classifier for this project to classify the histograms using linear kernel.

After we have trained the SVM, we do the evaluation and classification of each image in test class by mean Average Precision. Before we do the classification, we have to do normalized and standardized

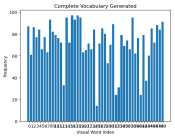


Figure 2: histogram of airplanes

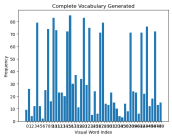


Figure 3: histogram of birds

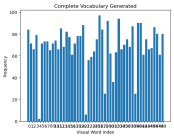


Figure 4: histogram of cars

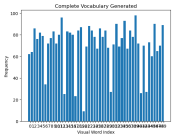


Figure 5: histogram of horses

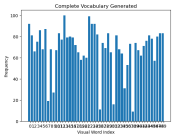


Figure 6: histogram of ships

Figure 7: histogram of different classes

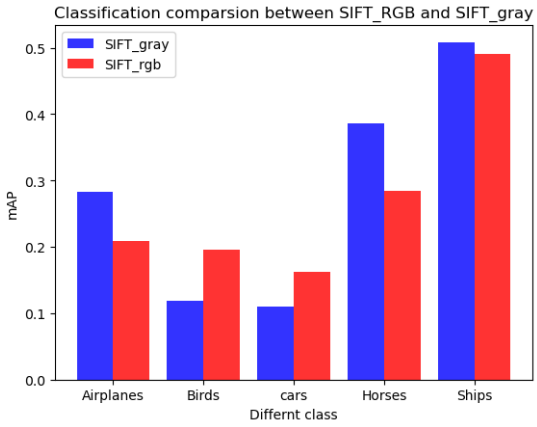


Figure 8: This is the mAP result comparing SIFT gray and sift rgb

in the same way of what we do extract descriptor in training class. Then we classify images using the histograms.

3 Experiments

Three experiments are performed, Firstly, we compare the classification result of SIFT and SIFT RGB algorithm with a vocabulary 300 words. RGB SIFT is one of the color sift descriptors, which gives a 384-dimensional descriptor that is formed from concatenating the 128-dimensional vectors from the three channels. So we do SIFT and extract the descriptors along each three channel, then then concatenate them. The result shows as Fig8. We can see that SIFT gray has a better classification result in horses and ships but less accuracy in airplanes and birds. Their classification quality in cars is similar.

In the second experiment, the classification effect of different vocabulary size is tested. We do the test of vocabulary size of(400 1000 4000), the configuration is SIFT gray descriptor. As result in Fig9. Different size of vocabulary seems has different accuracy in different class.The vocabulary size 400 do extremely a good job in airplane model but performs not good in other four class. On the contrary, the classification quality of 1000 and 4000 vocabulary is more average in different class. Hence, we can

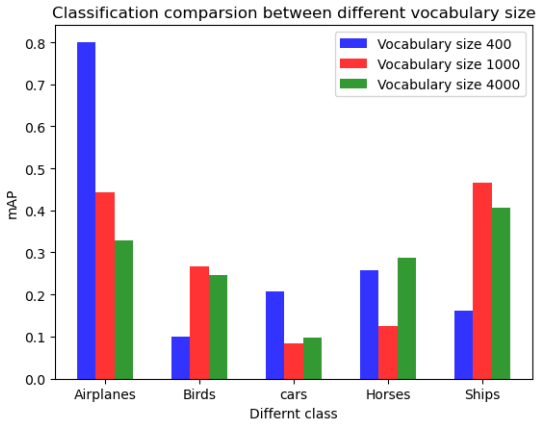


Figure 9: This is the mAP result comparing different vocabulary size

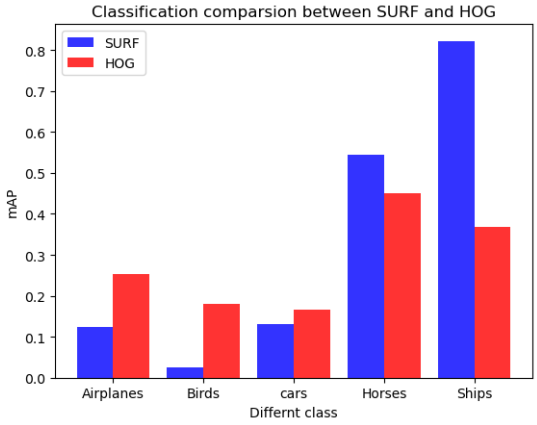


Figure 10: This is the mAP result comparing SURF and HOG

conclude that certain vocabulary size may have a good classification result in certain image class, so it is wise to choose proper vocabulary size base on what we need to classify.

In the third experiment, we compared the classification result between SURF and HOG result.The histogram of oriented gradients (HOG) and speeded up robust features (SURF) are another two feature descriptors used in computer vision beyond SIFT and image processing for the purpose of object detection.HOG Stands for histogram of oriented gradients basing on first order image gradients. The image gradients are pooled into overlapping orientation bins in a dense manner.SURF approximates the DoG with box filters. Instead of Gaussian averaging the image, squares are used for approximation since the convolution with square is much faster if the integral image is used. Including SIFT we talked about above they are all good descriptor method and they may have differences in speed and precision. As Fig10 shows, we did experiment on Surf and HOG, with 300 vocabulary size. We can find SURF has a better classification result in Horses and ships class, but HOG ac hive better result in Airplanes and Birds classification.

4 Conclusion

In a conclusion, we realize the whole Bags of words algorithm in from feature extraction and description, vocabulary building, features quantifying and images representing, classifier training and classification. We also tried different setting in descriptor(SIFT,SIFT RGB, SURF, HOG) and vocabulary(400,1000,4000) to find their influence on classification quality. The result is for training images we choose, SIFT RGB has a relative stable result. For future work, we want to study the which algorithm is best for certain class of image, so that we could use them in area like face recognition and automatic drive.

references

[1]HUANG, Chunde et HUANG, Jiaxiang. A Fast HOG Descriptor Using Lookup Table and Integral Image. arXiv preprint arXiv:1703.06256, 2017.

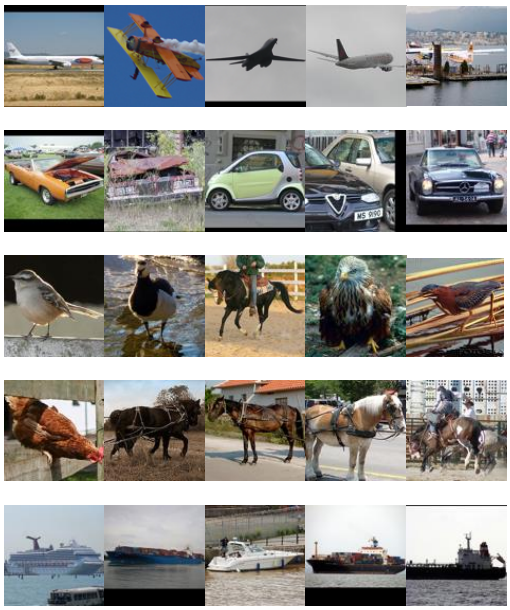
[2] YLIOINAS, Juha, KANNALA, Juho, HADID, Abdenour, et al.Face recognition using smoothed highdimensional representation. In : Scandinavian Conference on Image Analysis. Springer, Cham, 2015. p. 516-529.

[3] LU, Xiaojun, DUAN, Xu, MAO, Xiuping, et al. Feature Extraction and Fusion Using Deep Convolutional Neural Networks for Face Detection. Mathematical Problems in Engineering, 2017, vol. 2017.

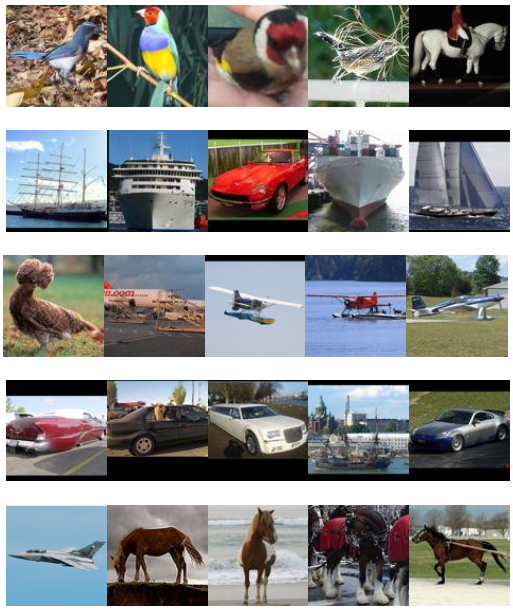
[4] M. EL BOUZ, F. BOUZIDI, A. ALFALOU, et al.," Adapted all-numerical correlator for face recognition applications. In : Optical Pattern Recognition XXIV," International Society for Optics and Photonics, 874807(2013).

[5] Q. WANG, A. ALFALOU, and C. BROSSEAU, "New perspectives in face correlation research: a tutorial,"Advances in Optics and Photonics 9,1-78 (2017).

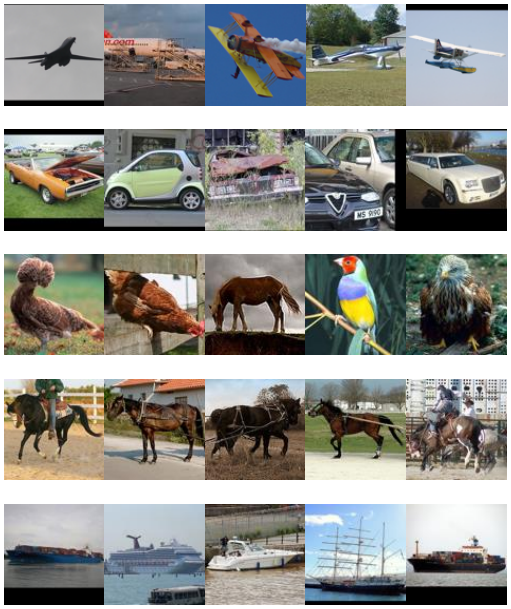
Appendix:



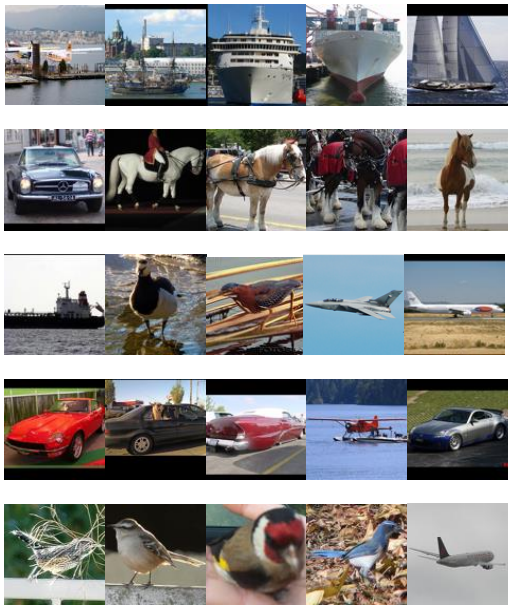
five-high ranked images for each class with SIFT 300 vocabulary size



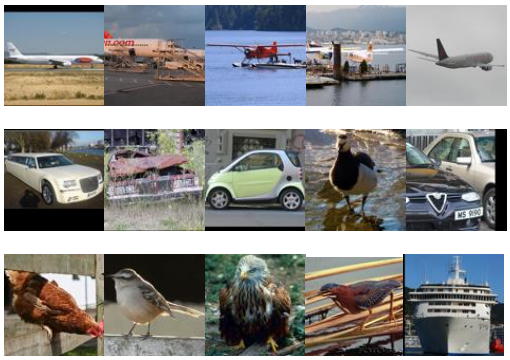
five-low ranked images for each class with SIFT 300 vocabulary size

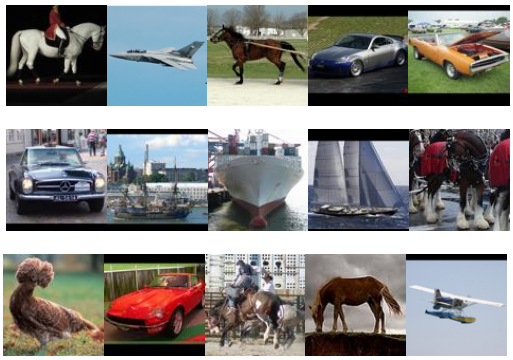


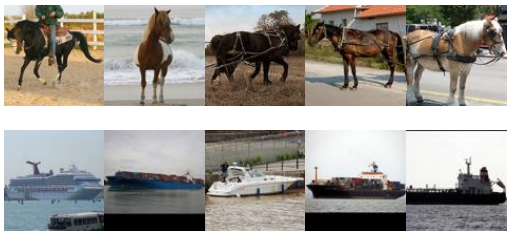
Five-high ranked images for each class with SIFT_RGB 300 vocabulary size



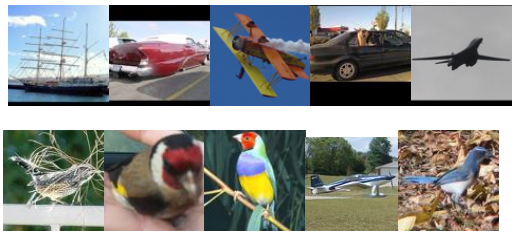
Five-low ranked images for each class with SIFT_RGB 300 vocabulary size



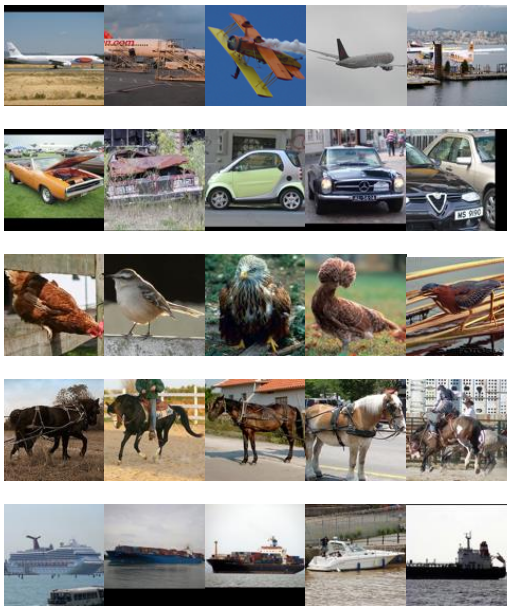




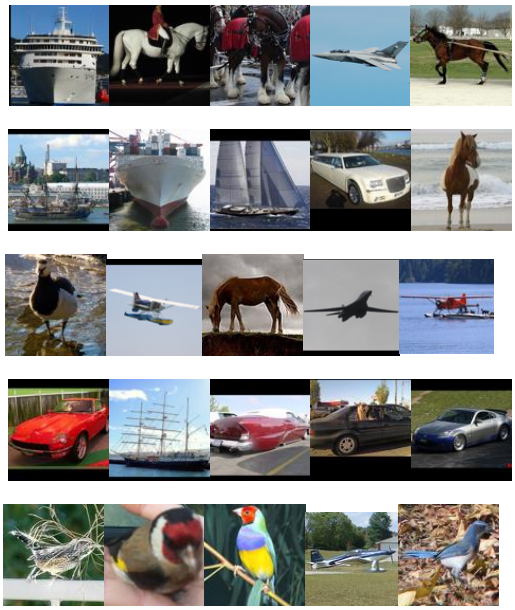
Five-high ranked images for each class with SIFT 1000 vocabulary size



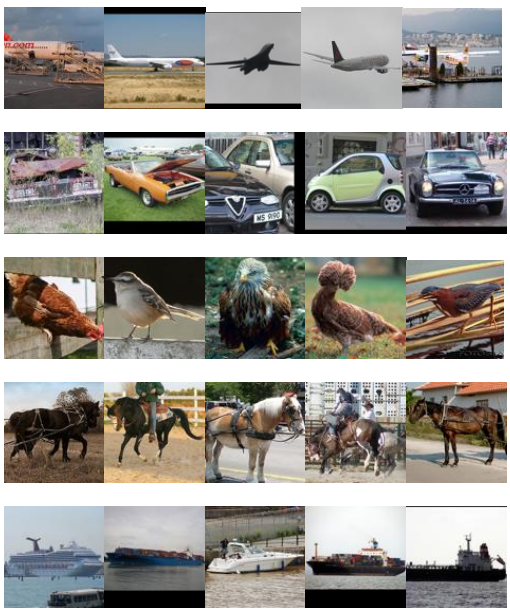
Five-low ranked images for each class with SIFT_RGB 1000 vocabulary size



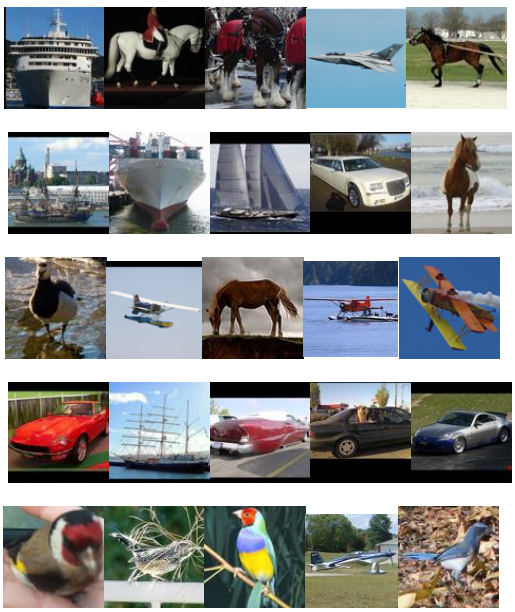
Five-high ranked images for each class with SIFT 4000 vocabulary size



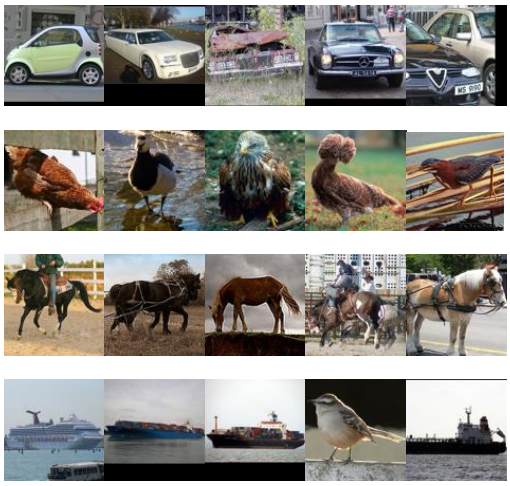
Five-low ranked images for each class with SIFT 4000 vocabulary size



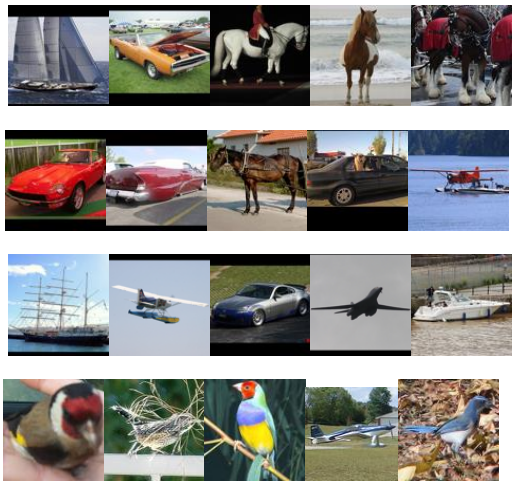
Five-high ranked images for each class with SURF 300 vocabulary size



Five-low ranked images for each class with SURF 300 vocabulary size



Five-high ranked images for each class with HOG 300 vocabulary size



Five-low ranked images for each class with HOG 300 vocabulary size